基于解耦表征学习的生成式视觉图像理解

蔡江海[1,2], 黄成泉[1,2,3)*], 王顺霞[2], 杨贵燕[2], 罗森艳[2], 周丽华[2]

[1] (贵州省模式识别与智能系统重点实验室 贵阳 550025)
[2] (贵州民族大学数据科学与信息工程学院 贵阳 550025)
[3] (贵州民族大学工程技术人才实践训练中心 贵阳 550025)
(hcq@gzmu.edu.cn)

摘 要: 学习可解释的视觉图像表征以揭示图像变化因素是计算机视觉领域的研究热点. 现有的许多解耦方法通过使用额外的正则项发现图像变化因素并学习解耦表征, 但通常导致解耦和生成质量之间的不平衡, 影响视觉图像理解效果. 为此, 从图像的可解释性变化出发, 提出基于解耦表征学习的生成式视觉图像理解方法. 首先设计预先训练的 Glow 生成模型, 获取目标图像的潜在表征; 然后由潜在表征构建基于图像变化的学习策略, 得到候选遍历的可解释方向; 最后在对比学习视角下设计对比模块, 根据候选遍历的可解释方向模拟图像变化, 进而提取解耦表征. 在解耦领域流行的数据集 Shapes3D, MPI3D, Anime, MNIST 和 Cars3D 上的实验结果表明, 所提方法取得较好的效果, 其中, 在 Cars3D 数据集上的 MIG, DCI, FactorVAE score 和 β-VAE score 指标值分别达到 0.16, 0.27, 0.89 和 0.98, 验证了该方法的有效性和可行性.

关键词: 解耦表征学习; 可解释方向; 图像生成; 对比学习
中图法分类号: TP391 **DOI:** 10.3724/SP.J.1089.2024-00003

Generative Visual Image Understanding Based on Disentangled Representation Learning

Cai Jianghai[1,2], Huang Chengquan[1,2,3)*], Wang Shunxia[2], Yang Guiyan[2], Luo Senyan[2], and Zhou Lihua[2]

[1] (*Key Laboratory of Pattern Recognition and Intelligent Systems of Guizhou Province, Guiyang* 550025)
[2] (*College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang* 550025)
[3] (*Engineering Training Center, Guizhou Minzu University, Guiyang* 550025)

Abstract: Interpretable visual image representation learning to reveal image variation factors is a hot research topic in computer vision. Many existing disentanglement methods discover variation factors of images and learn disentangled representations by using extra regularization term. However, it usually leads to an imbalance between disentanglement and generative quality, which affects visual image understanding. To address this issue, a generative visual image understanding method based on disentangled representation learning is proposed in terms of interpretable variations in images. Firstly, a pre-trained Glow model is designed to acquire the latent representations of target images. Secondly, a learning strategy based on image variation is constructed from the latent representations to obtain interpretable directions of candidate traversals. Finally, the contrast module is designed under the contrastive learning perspective to simulate image variations based on the interpretable directions of can-

收稿日期: 2024-01-02; 修回日期: 2024-07-12. **基金项目**: 国家自然科学基金(62062024); 贵州省科技计划(黔科合基础-ZK[2021]一般342); 贵州省研究生教育教学改革重点项目(黔教合 YJSJGKT[2021]018); 贵州省教育厅自然科学研究项目(黔教技[2022]015); 贵州省模式识别与智能系统重点实验室 2022 年度开放课题(GZMUKL[2022]KF03). **蔡江海**(1999—), 男, 硕士研究生, 主要研究方向为计算机视觉与模式识别、解耦表征学习; **黄成泉**(1976—), 男, 博士, 教授, 硕士生导师, 论文通信作者, 主要研究方向为计算机图形学、计算机应用技术、深度学习; **王顺霞**(1999—), 女, 硕士研究生, 主要研究方向为图像处理; **杨贵燕**(1997—), 女, 硕士研究生, 主要研究方向为图像处理; **罗森艳**(1999—), 女, 硕士研究生, 主要研究方向为图像处理; **周丽华**(1983—), 女, 硕士, 副教授, 主要研究方向为深度学习、图像处理.

didate traversals and then extract disentangled representations. The experimental results show that better results are achieved on the popular disentanglement datasets, which are Shapes3D, MPI3D, Anime, MNIST and Cars3D, where the MIG, DCI, FactorVAE score and β-VAE score metrics reach 0.16, 0.27, 0.89 and 0.98, respectively, on the Cars3D dataset, verifying the effectiveness and feasibility of the proposed method.

Key words: disentangled representation learning; interpretable direction; image generation; contrastive learning

解耦表征学习在人工智能领域有重要的研究价值, 旨在识别和学习隐藏在图像数据中的潜在可解释因素, 使深度神经网络像人类一样, 在观察事物关系时学会挖掘有意义的理解过程. 解耦表征学习作为一种表征范式, 常被定义为单个潜在维度的变化对应于单个生成因子的变化, 而相对其他因子的变化不敏感[1].

使用生成模型学习图像中的解耦表征具有一定的优越性. 当前, 许多方法在隐式或显式地了解生成式视觉图像真实潜在变化因子上表现出色. 变分自动编码器(variational auto-encoder, VAE)提供编码器和生成器[2], 生成对抗网络(generative adversarial network, GAN)提供生成器[3], 但在实际应用中, 这些穷尽的学习往往不可行. 在学习和推理过程中, 典型的方法还依赖于额外的正则项, 如β-VAE 和 FactorVAE 的总相关性[4], 或 Info-GAN (information maximizing-GAN)的互信息性[5]. 然而, 额外的正则项会导致解耦和生成质量之间的不平衡, 且当前许多无监督方法在不引入归纳偏差的情况下具有多个纠缠解, 不利于解耦表征学习. 此外, 在生成模型的潜在空间中, 沿着不同方向的遍历会使生成图像发生不同的变化, 表明生成模型的潜在空间具有一定解耦特性.

对于生成模型, VAE 的目标是最大化目标图像数据对数似然的变分下界, 且采用 KL(Kullback-Leibler)散度进行分布衡量[2]. GAN 使用生成器和判别器进行博弈, 找到其中的纳什平衡点[3]. 在精确的潜在变量推理和对数似然估计上, VAE 仅粗略地推断出图像数据点对应的潜在变量值; 在 GAN 中, 图像数据通常不能直接在潜在空间中表示. 与这 2 种模型不同, Glow(generative flow)建立一系列可逆变换并直接优化数据分布的对数似然, 其具有如下优点: (1) 利用可逆生成模型不仅能实现准确的推理, 还可以优化图像数据的精确对数似然; (2) 在并行硬件上具有高效的并行推理和综合处理能力; (3) 允许对图像数据进行有意义的修正, 获得对下游任务有意义的潜在空间; (4) 由于模型的结构特点, 其具有节省内存的潜力. 因此, 使用 Glow 能更好地获得目标图像的潜在表征, 为解耦表征学习任务奠定基础.

为了解决上述问题, 本文提出基于解耦表征学习的生成式视觉图像理解(disentanglement for generative visual image understanding, Dis-GU)方法. 该方法固定可逆的基于流的生成模型 Glow, 联合学习生成模型潜在空间中的可解释方向, 并构建变化空间, 进而提取解耦表征. 其中, 通过在潜在空间中遍历方向, 实现潜在变化因子的发现; 采用一个典型的网络模块(学习器), 提供潜在空间中的候选遍历的可解释方向. 在解耦表征学习中, 通过对比学习构建变化空间, 模拟图像的各种变化; 并在此空间中提出关键的重建损失技术, 以优化模型. 与多种典型的解耦方法进行实验的结果表明, Dis-GU 的解耦效果显著.

1　相关工作

1.1　图像表征的解耦与生成

发现并学习图像的潜在变化因子是解耦表征学习研究的目标之一. 在图像表征的解耦与生成研究中, 许多工作集中在解耦图像属性和区分风格内容上[6]. Chen 等[4]提出通过最大化 GAN 进行可解释性的表征学习; Chen 等[7]利用 GAN 的信息论扩展学习复杂环境的解耦表征; Xiao 等[8]从多属性图像出发学习潜在变化因子; Higgins 等[9]提出 β-VAE, 在变分先验和变分后验的 KL 散度上使用更大的权重 $(\beta>1)$, 这是一种有效且稳定的解耦方法; Kim 等[1]提出 FactorVAE, 鼓励边际分布是阶乘的, 并使用密度比技巧最小化 KL 散度, 以获得更好的解耦效果: 这些方法使用额外的正则项进行解耦表征学习, 虽然具有一定解耦特性, 但往往牺牲重建质量. 本文通过对生成图像的潜在空间进行语义上有意义的探索, 并考虑单个潜在变化因子对应同类图像的一个属性特征变化, 在解耦和重建质量之间获得更好的平衡, 实现多尺度解耦[10].

1.2 潜在空间中的可解释方向

目前，许多研究聚焦于生成模型潜在空间中的可解释方向. 生成模型的潜在空间通常具有语义上的可解释性，许多方法以此进行视觉图像的相关任务，如图像重建、图像编辑和解耦图像特征等. Shen 等[11]利用伪标签在潜在空间中构造分离超平面，以捕获相应属性特征的语义方向，并通过计算语义分数的均值和协方差建立解耦表征的联系；Dalva 等[12]在潜在空间中设计属性编辑，并对每个属性学习一个与其他属性正交的线性方向；Endo[13]通过注释来解决图像的布局问题，并估计输出潜在编码获得可解释方向；Xu 等[14]通过构建相互访问约束和反演一致性约束，并联合优化基准图像的潜在编码和语义方向提高潜在编码的可编辑性. 然而，这些方法只能发现目标期望中的可解释方向，且在解耦表征效果上表现较差，而在实际应用中，这些方向远不能对图像数据进行有效的表征. 相比之下，Dis-GU 从生成模型的潜在空间出发，识别出更丰富的可解释方向，为后续解耦表征学习的下游任务提供更有力的支撑.

1.3 对比学习

随着人工智能技术的发展，对比学习在解耦领域应用广泛. 当前的工作将对比学习应用于生成模型，在解耦表征效果上具有一定竞争力[15]. 对比学习常被扩展到各种下游任务，如小样本学习、可控生成和图像翻译. 通常，对比学习方法使同类图像(正对)不同视角的表征更加接近，并使不同图像(负对)视角的表征分开[16]. 将对比学习应用于解耦表征学习出于以下考虑：(1) 在实际应用中，解耦方向的实际数目是未知的，通过对比学习可类似于检索，能够发现有效的解耦方向[17]；(2) 对比学习可直接在变化空间中工作，无需任何额外的约束[18]. 因此，本文将对比学习应用于表征变化，使之关注变化空间中的解耦样本，获得对视觉图像的理解.

2 Dis-GU

2.1 任务描述

在解耦表征学习中，生成式视觉图像理解的本质是将图像中混合一系列物理概念的潜在变化因子$\{z_0, z_1, \cdots, z_n\}$进行识别和学习，并提取解耦表征，如图 1 所示. 此外，希望在改变一个潜在变化因子时仅生成关于该单因子的变化图像，且该图像的其他属性不变. 当前，许多传统的解耦方法不能达到理想的效果：在改变某一潜在变化因子进行图像生成时，存在至少 2 个物理属性纠缠在一起的问题(未实现单因子解耦). 如图 2 所示，生成的图像中将地板颜色属性和墙壁颜色属性纠缠一起. 这已成为目前生成式视觉图像理解领域亟需解决的问题.

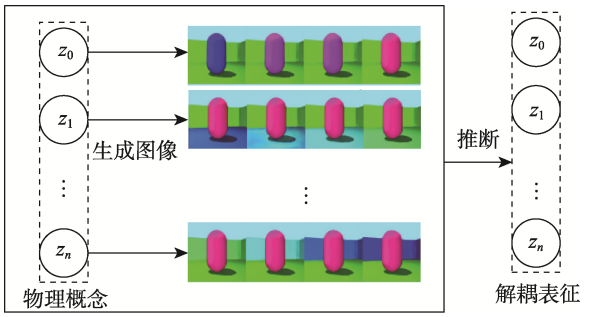

图 1 解耦表征的过程

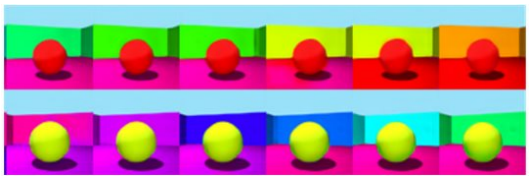

图 2 2 个属性纠缠的示例

2.2 概 览

从解耦表征学习的直观概念出发，Dis-GU 利用预先训练的 Glow 模型和学习策略，联合发现嵌入在生成模型潜在空间中的可解释方向，并利用对比模块构建变化空间，以提取解耦表征，实现对视觉图像的解耦表征学习.

Dis-GU 由 2 部分组成：(1) 可解释方向的发现阶段. 提供视觉图像潜在空间中的候选遍历的可解释方向；(2) 解耦表征学习阶段. 对已发现的方向进行有意义的图像变化表征提取. 其中，Glow 模型用于训练目标图像数据集，使潜在空间中的潜在编码能够刻画目标图像的数据分布；学习器用于发现潜在空间中的候选遍历的可解释方向；解耦编码器用于模拟图像的各种变化，并将 2 幅图像之间的变化建模为其提取对应编码表示的差值. Dis-GU 的结构如图 3 所示.

在解耦编码器的作用下，同一方向的解耦样本集中为一类，不同方向的解耦样本相互区分. 解耦编码器将图像编码到变化空间，其中应用了对比学习的思想；变化空间中的解耦样本对应于已发现的可解释方向的图像变化. 因此，相同因子的

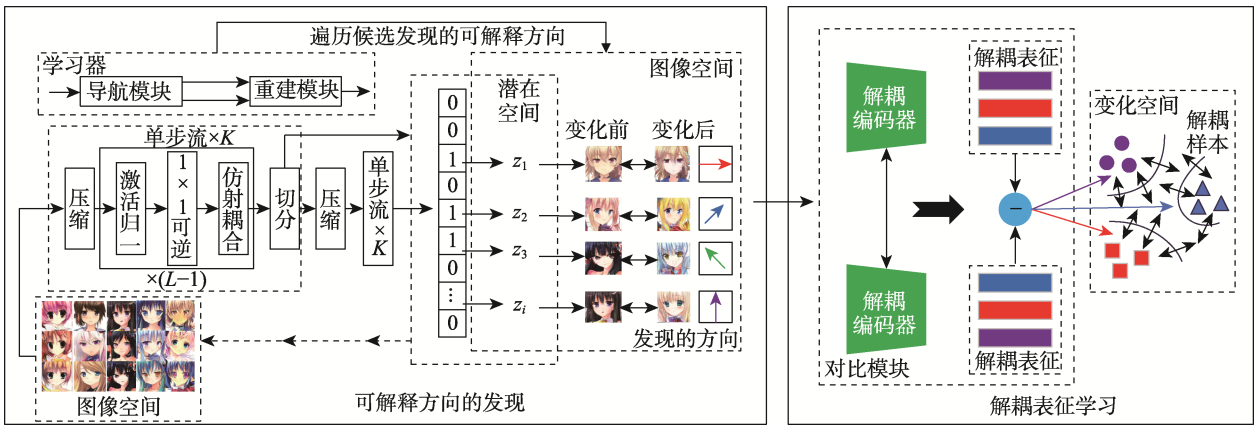

图 3　Dis-GU 的结构

变化对应于相似的图像变化，不同因子的变化对应于不同的图像变化，为解耦表征学习提供了新的对比学习视角.

2.3　工作流程

给定目标生成式视觉图像数据和预先训练的生成模型 Glow：$\{G:Z\to I\}$，其中，Z 表示潜在空间，I 表示图像空间. Dis-GU 的步骤如下：

Step1. 通过 Glow 模型训练目标图像，获得潜在表征(编码表示 $[0,0,1,\cdots,0,1,\cdots,0]$，其中，1 表示发现的潜在编码表示).

Step2. 在潜在空间中利用学习器 Ω_{LD} 遍历发现的方向，得到候选的可解释方向 $\{z_1,z_2,z_3,\cdots,z_i\}$；其中，$z_i$ 表示候选的潜在编码.

Step3. 将候选的可解释方向映射到图像空间，生成图像对 $\left(G(z),G\left(z+N(\varepsilon\boldsymbol{\gamma}_{\mathrm{b}})\right)\right)$；其中，$z$ 表示潜在编码，$\boldsymbol{\gamma}_{\mathrm{b}}$ 表示单位向量 $(0,\cdots,1_{\mathrm{b}},\cdots,0)$，$\varepsilon$ 表示位移大小，N 表示导航模块，$G(z)$ 表示从 Glow 模型的潜在编码 z 生成第 1 幅图像，$G\left(z+N(\varepsilon\boldsymbol{\gamma}_{\mathrm{b}})\right)$ 表示从位移编码 $z+N(\varepsilon\boldsymbol{\gamma}_{\mathrm{b}})$ 生成第 2 幅图像.

Step4. 将可解释方向的发现阶段获得的图像对 $\left(G(z),G\left(z+N(\varepsilon\boldsymbol{\gamma}_{\mathrm{b}})\right)\right)$ 通过对比模块 Θ_{E} 编码为解耦样本 $v\in V_{\mathrm{CL}}$：$v=\left|E\left(G\left(z+N(\varepsilon\boldsymbol{\gamma}_{\mathrm{b}})\right)\right)-E(G(z))\right|$；其中，$V_{\mathrm{CL}}$ 表示变化空间；E 表示解耦编码器. 利用 Θ_{E} 使解耦表征达到分离维度的响应，在变化空间中基于解耦样本提取解耦表征，提升生成式视觉图像理解效果.

2.4　潜在表征的获取

根据 Glow 模型的优良特性[19]，本文使用其进行目标图像训练，获取目标图像的潜在表征. 在 Glow 模型中的生成过程为

$$z\sim p_\theta(z) \tag{1}$$

$$x=F_\theta(z) \tag{2}$$

其中，z 表示潜在变量；x 表示训练数据；$p_\theta(z)$ 表示可解的概率密度函数. 从简单的球形多元高斯分布 $p_\theta(z)=N(z;0,I)$ 采样得到 z，$F_\theta(z)$ 是可逆的(双射)，有 $z=f_\theta(x)=F_\theta^{-1}(x)$. 此外，使用一系列的变换来学习图像的生成和推理过程，通过可逆变换的训练，逐步获得目标图像的数据分布，以表示目标图像数据的有效表征(存在于潜在空间). Glow 模型的训练生成和推理过程如图 4 所示.

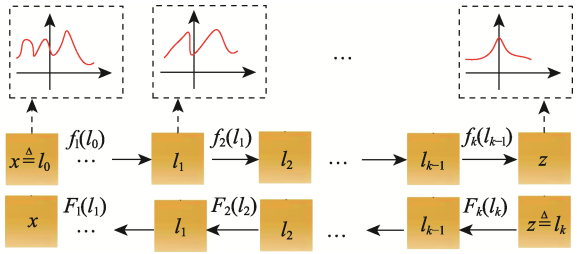

图 4　Glow 模型的训练生成和推理过程

为了提高潜在表征的有效性，使用的训练优化目标函数为

$$\lg p_\theta(x)=\lg p_\theta(z)+\lg\left|\det\left(\frac{\mathrm{d}z}{\mathrm{d}x}\right)\right|=\lg p_\theta(z)+\sum_{i=1}^{K}\lg\left|\det\left(\frac{\mathrm{d}l_i}{\mathrm{d}l_{i-1}}\right)\right| \tag{3}$$

其中，$\lg\left|\det\left(\dfrac{\mathrm{d}l_i}{\mathrm{d}l_{i-1}}\right)\right|$ 表示从 l_{i-1} 到 l_i 目标函数变化过程中的改变量，可以使用雅可比矩阵为三角矩阵的变换进行计算，其对数行列式存在关系

$$\lg\left|\det\left(\frac{\mathrm{d}l_i}{\mathrm{d}l_{i-1}}\right)\right|=s\left(\lg\left|\mathrm{diag}\left(\frac{\mathrm{d}l_i}{\mathrm{d}l_{i-1}}\right)\right|\right) \tag{4}$$

其中，$s(\cdot)$ 表示对相关元素求和；$\mathrm{diag}(\cdot)$ 表示获取的雅可比矩阵的对角元素.

Glow 模型每步的变换训练过程由激活归一操作层、1×1 可逆卷积层和仿射耦合层 3 个串行网络

模块构成，该模型详见文献[19]，其单步流结构和多尺度训练结构如图 5 所示.

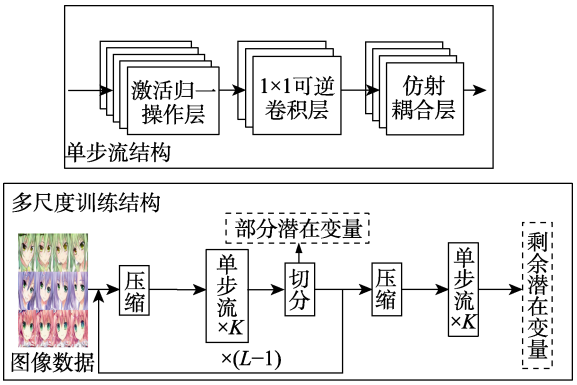

图 5 Glow 模型的单步流结构和多尺度训练结构

2.5 可解释方向的发现

为了发现生成模型潜在空间中的可解释方向，学习一组引发“正交”图像变换的方向是有意义的[20]. 本文参考生成模型潜在空间中可解释方向的相关研究[20]，构建用于视觉图像理解的学习器 Ω_{LD}，目标是发现经过 Glow 模型训练后的视觉图像的可解释方向，并将获得的方向从潜在空间映射到图像空间 $G: Z \to I$. 其中，Glow 模型在该阶段为不可训练组件，其参数在学习过程中不会改变.

Ω_{LD} 的 2 个可训练组件分别为：(1) 导航模块 N. 该模块由矩阵 $\boldsymbol{M} \in \mathbb{R}^{a \times b}$ 构成，其中，a 表示 Glow 模型潜在空间的维数，b 表示发现的方向数. $\boldsymbol{M}$ 为 Dis-GU 的一个超参数，本质上，其列数对应期望识别的方向数. 由于空间限制，本文仅考虑实现效果良好的 $\boldsymbol{M}$ 的列(具有单位长度和具有正交矩阵列的线性算子). (2) 重建模块 R. 可获得图像对 $\left(G(z), G\left(z+N\left(\varepsilon \boldsymbol{\gamma}_{\mathrm{b}}\right)\right)\right)$，其中，第 1 幅图像从 Glow 模型的潜在编码 z 生成，而第 2 幅图像从位移编码 $z+N\left(\varepsilon \boldsymbol{\gamma}_{\mathrm{b}}\right)$ 生成. 注意，第 2 幅图像为第 1 幅图像的变换，对应于在 $\boldsymbol{M}$ 定义的方向上(第 b 列)移动 ε 大小. R 的目标是重建潜在空间中引起给定图像变换的方向. 理论上，R 执行映射 $R\{I_1, I_2\} \to \left(\{1,2,\cdots,b\}, \mathbb{R}\right)$，且产生 2 个输出 $R\{I_1, I_2\} = \left(\hat{d}, \hat{\varepsilon}\right)$，其中，$\hat{d}$ 和 $\hat{\varepsilon}$ 分别表示方向索引 $d \in \{1,2,\cdots,b\}$ 和偏移 ε 大小的预测. Ω_{LD} 的优化目标函数为

$$\min_{N,R} \mathop{E}_{z,d,\varepsilon} L(N,R) = \min_{N,R} \mathop{E}_{z,d,\varepsilon} \left[L_{\mathrm{c}}\left(d, \hat{d}\right) + \eta L_{\mathrm{r}}\left(\varepsilon, \hat{\varepsilon}\right) \right] \quad (5)$$

其中，$\mathop{E}_{z,d,\varepsilon}$ 表示获取 z，d 和 ε 条件下的最优值；分类项 $L_{\mathrm{c}}(\cdot,\cdot)$ 使用交叉熵函数；回归项 $L_{\mathrm{r}}(\cdot,\cdot)$ 使用平均绝对误差；在本文的实验中，超参数 $\eta = 0.25$.

Ω_{LD} 中的导航模块 N 和重建模块 R 是联合优化的，其在最小化优化目标函数的过程中寻求获得 N 中具有解释性的列数 b，使相应图像变换更容易区分，从而在可解释方向的分类问题上更简单. Ω_{LD} 发现可解释方向的过程如图 6 所示.

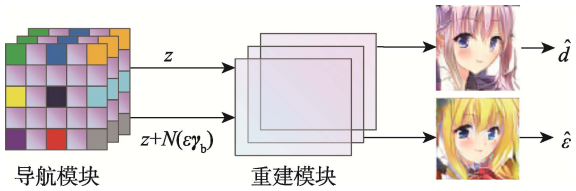

图 6 学习器发现可解释方向的过程

2.6 解耦表征学习

本文应用对比学习的思想，将学习器 Ω_{LD} 中发现的可解释方向进行解耦表征学习. 设计基于解耦编码器的对比模块 Θ_{E}，构建变化空间 V_{CL}，并将图像对 $\left(G(z), G\left(z+N\left(\varepsilon \boldsymbol{\gamma}_{\mathrm{b}}\right)\right)\right)$ 编码为解耦样本，以提取解耦表征[18]. Θ_{E} 将 2 幅图像之间的变化建模为由解耦编码器提取的对应编码表示的差异，且沿着发现的方向移动会引起明显的图像变化，使解耦表征达到分离维度的响应. 因此，在 V_{CL} 中将遍历同一方向的解耦样本拉在一起，并将遍历不同方向的解耦样本推离，提升生成式视觉图像理解效果. Θ_{E} 的结构如图 7 所示.

二元交叉熵(binary cross entropy, BCE)损失函数[18]已被广泛应用于实现对比学习中. 为了使解耦表征学习的结果更好，本文提出基于 BCE 损失函数的重建损失 L_{R}，其遵循文献[21]的相关理论，是 BCE 损失函数的推广. 存在配对观察 $x_i: (q, k_i)$，q 表示查询键，键 k_i 来自正键集 $\left\{k_j^+\right\}_{j=1}^N$ 或负键集 $\left\{k_m^-\right\}_{m=1}^M$，即 $\left\{k_i\right\}_{i=1}^{N+M} = \left\{k_j^+\right\}_{j=1}^N \cup \left\{k_m^-\right\}_{m=1}^M$，配置项为

$$C_i = \begin{cases} 1, & k_i \in \left\{k_j^+\right\}_{j=1}^N \\ 0, & k_i \in \left\{k_m^-\right\}_{m=1}^M \end{cases} \quad (6)$$

根据 BCE 损失函数[18]，重建损失 L_{R} 的一般形式为

$$L_{\mathrm{R}} = -\sum_{i=1}^{S} C_i \lg \sigma\left(q \cdot k_i / \tau\right) + \left(1 - C_i\right) \lg\left(1 - \sigma\left(q \cdot k_i / \tau\right)\right) \quad (7)$$

其中，$\sigma(\cdot)$ 表示 Sigmoid 函数；τ 表示超参数；S

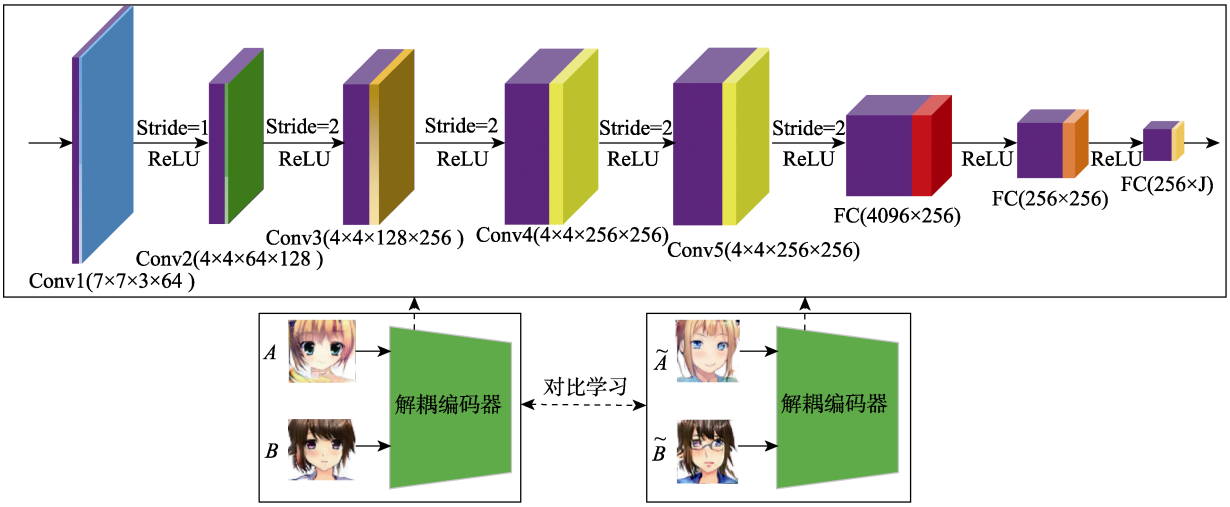

图 7　对比模块的结构

表示样本数. 推广到本文的设置(N 个正样本和 M 个负样本), 式(7)重新表述为

$$\hat{L}_{\mathrm{R}}=-\sum_{j=1}^{N}\lg\sigma\left(q\cdot k_j^{+}/\tau\right)-\sum_{m=1}^{M}\lg\left(1-\sigma\left(q\cdot k_m^{-}/\tau\right)\right) \quad (8)$$

根据式(6)~(8), 令 $h\left(x_i\right)=\exp\left(q\cdot k_i/\tau\right)$, 其在对比学习中有意义. 通过正负 2 个方面的对比, 将 $h\left(x_i\right)$ 作为相似度得分[18]. 如果 q 和 k_i 来自一个正对(具有相同的可解释方向), 那么 $h\left(x_i\right)$ 得分应尽可能大 $\left(h\left(x_i\right)>1\right)$; 反之亦然. 因此, Dis-GU 可学习反映图像变化的表征 $\left(q,k_i\right)$, 即图像的相似变化具有较高的 $h\left(x_i\right)$ 得分, 而图像的不同变化将具有较低的 $h\left(x_i\right)$ 得分. 此外, 结合表征学习和聚类任务的理解[22], 利用这种有意义的表示发现不同类型图像变化的可解释方向, 提取解耦表征.

3　实验及结果分析

为了全面地评估 Dis-GU 的解耦性能, 分别进行 5 个模块的实验: Dis-GU 直观分析、Dis-GU 解耦探索、定性分析、定量分析和参数分析.

3.1　实验设置

3.1.1　实施细节

本文实验中, 关键参数设置如下: 学习率为 10^{-4}, 批量大小为 32, 负样本数为 64, 方向数为 64, 阈值为 0.95; 使用 Adam 优化器, 并基于交叉熵损失进行训练优化, 训练迭代次数达 70 000 次, 以每 10 000 次迭代为一个训练节点. 为了减少随机性的干扰, 所有训练保证运行 10 次以上. 部分对比方法的超参数设置遵循当前最佳的结果设置. 对于 β-VAE, 设置 β=5; 对于 FactorVAE, 设置 γ=9.

3.1.2　数据集

使用解耦研究中流行的数据集: (1) Shapes3D[1]. 包含 48 000 幅 RGB 图像的三维合成数据集. (2) Cars-3D[23]. 使用 199 个 CAD 模型, 从 24 个旋转角度合成的三维渲染图像数据集. (3) MPI3D[24]. 具有 7 个潜在变化因子, 且由三维物体图像组成的数据集. (4) MNIST[25]. 由像素级黑白图像组成的手写数字数据集. (5) Anime[21]. 包含不同大小的动漫图像数据集.

3.1.3　评估指标与基线

为了从多个角度衡量 Dis-GU 的解耦性能, 使用在解耦研究中应用广泛的评估指标[1]: 互信息差异(mutual information gap, MIG), 用于度量每个潜在变化因子的信息被潜在编码的单个维度捕获的程度; 解耦性/完整性/信息性(disentanglement, completeness and informativeness, DCI), 用于度量信息生成嵌入的解耦性、完整性和信息性; FactorVAE score, 用于度量数据潜在表征的因子尺度; β-VAE score, 用于度量潜在变化因子的独立性和解释性. 从定性基准和定量基准 2 个层面, 将 Dis-GU 与典型的解耦基线进行比较: β-VAE[9], FactorVAE[1], DIP-VAE(disentangled inferred prior-VAE)[26], CF(closed form)[27], GS(GAN space)[27], DS(deep spectral)[27], LD(latent discovery)[27]和 DisCo (disentanglement via contrast)[18].

3.2　Dis-GU 直观分析

在解耦表征学习中, 应使模型学会发现具有明显变化模式的可解释方向, 并排除随机方向. 本节以 Shapes3D 数据集为例, 直观地分析 Dis-GU 发现可解释方向, 如图 8 所示.

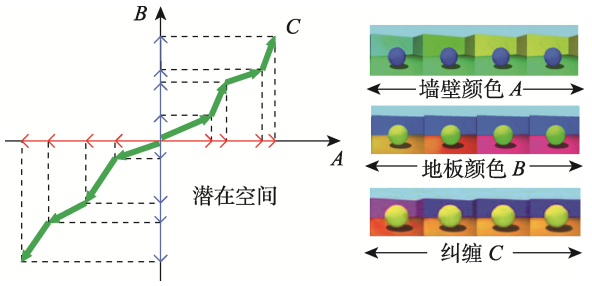

图 8 Dis-GU 发现可解释方向

假设在生成模型的潜在空间中存在因子 A 和因子 B 的解耦方向，因子 C 的方向为因子 A 和因子 B 的纠缠方向(随 A 和 B 一起变化，包含 A 和 B 的信息)，Dis-GU 的目标是发现可解释方向(A 和 B)，并排除随机方向(C). 从图 8 可以看出，Dis-GU 发现了因子 A 和因子 B 的可解释方向，对于因子 C，由于 Dis-GU 没有额外的偏差来识别 C 的方向，因此 C 的随机方向被排除. Dis-GU 几乎不会收敛到 A 和 C 同时存在的情形，原因如下: (1) Dis-GU 的编码器是一个轻量化网络，几乎不能将 A，B 和 C 的方向区分开; (2) A 和 B 的解耦方向在不同位置正交，而 C 中的方向不一致(包含至少 A 和 B 融合的信息)，且其变化的“速率”不同，因此存在纠缠的结果(如将墙壁颜色和地板颜色纠缠在一起). 综合解耦方向和纠缠方向在潜在空间中的不同性质可知，其解耦方向的变化比纠缠方向的变化更一致，并且能够更好地聚类. 因此，Dis-GU 能够较好地识别具有明显变化模式的可解释方向，并从图像中学习解耦表征.

3.3 Dis-GU 解耦探索

为了进一步验证 Dis-GU 能够发现潜在空间中的解耦方向且不收敛于纠缠的方向，以 Shapes3D 数据集为例进行实验. 选择地板颜色和背景颜色 2 个地面真值因子，沿着解耦的方向进行等间距采样，获得地板颜色因子、背景颜色因子和复合因子的解耦方向，如图 9a 所示; 另一方面，通过解耦表征的结果和经验设置，将 Dis-GU 在 Shapes3D 数据集中的生成图像回归为地面真值因子(地板颜色因子和背景颜色因子)的解耦表征，如图 9b 所示.

从图 9b 可以看出，表示地板颜色因子的样本(①)与水平轴较好地对齐，表示背景颜色因子的样本(②)与垂直轴也较好地对齐，但是复合样本(③)未与任何一条轴线对齐，且其样本分布也较为分散，表明复合因子的方向不一致. 实验结果表明，Dis-GU 能够发现解耦的方向，且能够较好地识别并区分纠缠的方向，进一步验证了 Dis-GU 能够进行有效的解耦表征学习.

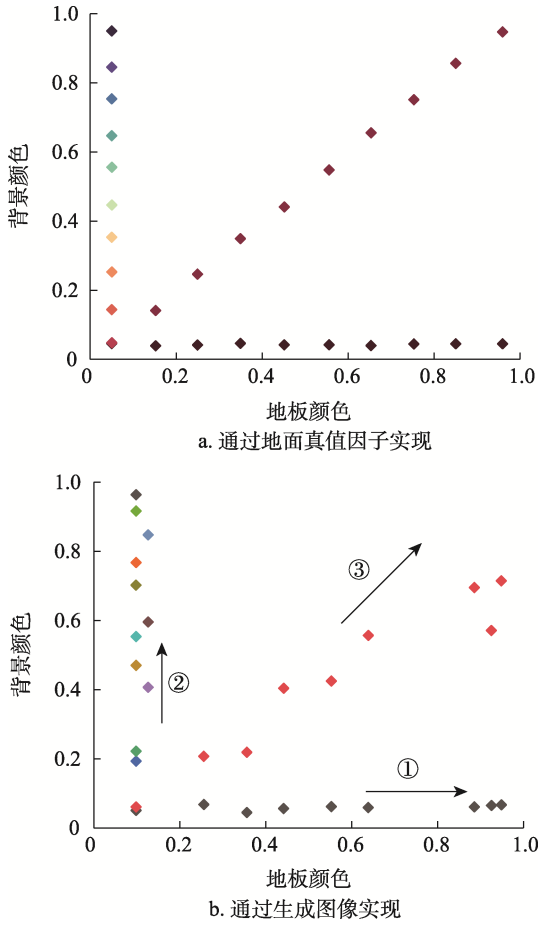

图 9 潜在空间中的解耦方向

3.4 定性分析

在 Shapes3D, Anime, MNIST, Cars3D 和 MPI3D 这 5 个数据集上对 Dis-GU 进行定性实验，结果如图 10~图 14 所示. 可以看出，Dis-GU 在 5 个数据集上均获得了理想的结果，即当一个潜在维度发生变化时，其仅对应于单个生成因子的变化，而相对其他因子的变化不敏感(生成的图像不存在多个属性纠缠的现象). 图 10 中，分别发现地板颜色因子(上)、物体颜色因子(中)和墙壁颜色因子(下)，学习更清晰和独立的表征，实现了单因子解耦; 图 11 中，解耦人物图像的头发颜色因子(上)和眼镜因子(下)，获得更加丰富的图像生成样式，为特定任务提供了参考样本; 图 12 中，对于手写数字，分别发现数字厚度因子(上)、数字角度因子(中)和数字宽度因子(下)，倾向于更宽且平滑的连续变换学习，进而获得更为丰富的因子; 图 13 中，分别发现车颜色因子(上)和车样式因子(下)，这些因子可解释方向的理解在实际应用中效果显著; 图 14 中，对于抽象体，分别发现抽象体的旋转角度因子(上)和颜色因子(下).

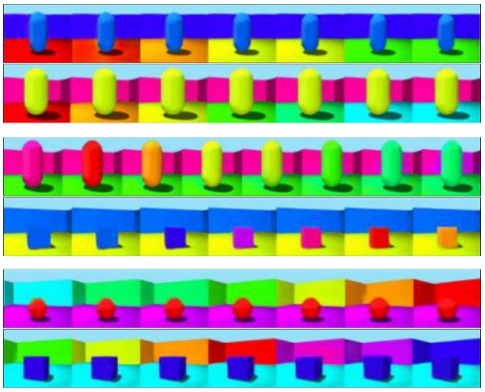

图 10　在 Shapes3D[1]上可解释方向的发现

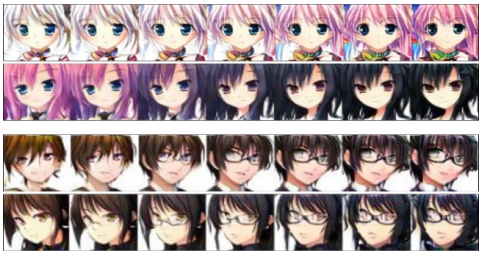

图 11　在 Anime[21]上可解释方向的发现

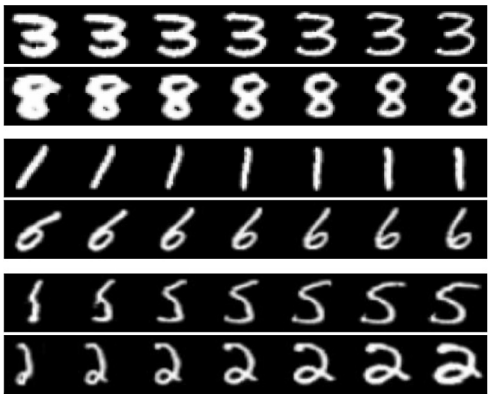

图 12　在 MNIST[25]上可解释方向的发现

图 13　在 Cars3D[23]上可解释方向的发现

总体上，Dis-GU 能够发现更多的潜在变化因子，且发现的方向易于解释；此外，还学习更干净、更独立的表达方式，生成具有单因子解耦属性的图像. 定性实验结果表明，在生成式视觉图像理解上，Dis-GU 的解耦表征学习效果较好.

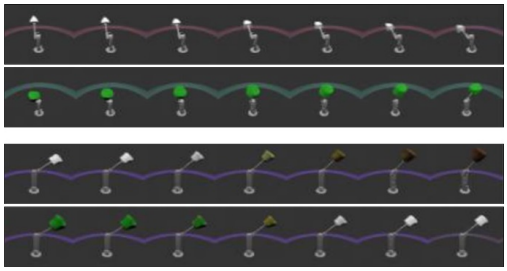

图 14　在 MPI3D[24]上可解释方向的发现

3.5　定量分析

在 Shapes3D，MPI3D 和 Cars3D 这 3 个数据集上将 Dis-GU 与解耦领域典型的方法进行实验，定量地分析该方法的解耦性能，结果如表 1 所示，其中，解耦评分越高，解耦性能越佳；第 3~6 列中，结果表示形式为均值±标准差. 可以看出，(1) Dis-GU 获得了具有竞争力的结果，其解耦性能较优. (2) 尤其在 Shapes3D 上，Dis-GU 在解耦性能区分度上效果显著，其中，比 DisCo 方法在解耦评分上分别高出 0.30，0.06，0.07 和 0.12；在 Shapes3D 和 MPI3D 上，可能由于随机种子选取的影响，CF 和 DS 方法的 FactorVAE score 和 β-VAE score 这 2 个指标的解耦评分较高. (3) Dis-GU 还获得了较小的方差，表明模型的训练相对稳定. 实验结果表明，Dis-GU 在不同体系结构、参数优化和数据集上具有一定鲁棒性，能够稳定地发现更多的潜在变化因子，且学习的解耦表征涵盖更广泛的因子值范围.

3.6　参数分析

为了进一步验证 Dis-GU 的解耦性能，在 Shapes3D 数据集上分别对 Dis-GU 中的重要参数进行分析.

(1) 负样本数 K. 在 Dis-GU 中，K 的选取对模型的训练尤为重要. 引入负样本一是为了减少模型训练的计算量；二是为了提升模型的效果，使之具有更高的辨识图像组和捕捉细节特征的能力. 图 15 所示为对 K 进行分析的结果(K 的取值分别为 8, 16, 32, 64 和 128). 可以看出，Dis-GU 的解耦性能呈先增后减的趋势，当 K=64 时，Dis-GU 的解耦性能最佳. 因此，在实验中设置超参数 K=64.

(2) 方向数 D. 在可解释方向的发现阶段，若设定过多的方向数，则模型可能会捕捉过多纠缠方向的因子；否则，可能会忽视关键的解耦因子，使解耦表征学习和图像生成效果较差. 图 16 所示为对 D 进行分析的结果(D 的取值分别为 8, 16, 32, 64, 128 和 256). 可以看出，随着 D 的增加，Dis-GU 的推理成本也相应增加，当 D=64 时，Dis-GU 的

表 1　在 3 个数据集上不同方法的解耦评分对比

数据集	方法	MIG	DCI	FactorVAE score	β-VAE score
Shapes3D[1]	β-VAE[9]	0.40±0.12	**0.61±0.11**	0.87±0.07	0.93±0.03
	FactorVAE[1]	0.43±0.14	0.60±0.10	0.86±0.06	0.87±0.04
	DIP-VAE[26]	0.38±0.15	0.57±0.14	0.68±0.06	0.83±0.04
	CF[27]	0.30±0.16	0.51±0.08	**0.95±0.02**	0.98±0.06
	GS[27]	0.11±0.03	0.28±0.04	0.80±0.09	0.93±0.05
	DS[27]	0.36±0.07	0.51±0.07	0.93±0.02	0.97±0.02
	LD[27]	0.17±0.05	0.38±0.06	0.44±0.19	0.60±0.20
	DisCo[18]	0.15±0.00	052±0.00	0.85±0.00	0.86±0.00
	Dis-GU	**0.45±0.04**	0.58±0.03	0.92±0.06	**0.98±0.01**
MPI3D[24]	β-VAE[9]	0.10±0.05	0.24±0.06	0.18±0.01	0.35±0.01
	FactorVAE[1]	0.11±0.02	0.26±0.05	0.16±0.02	0.34±0.02
	DIP-VAE[26]	**0.13±0.06**	0.25±0.08	0.22±0.04	0.32±0.10
	CF[27]	0.08±0.02	0.29±0.02	0.51±0.06	0.61±0.03
	GS[27]	0.12±0.07	0.23±0.03	0.47±0.03	0.59±0.06
	DS[27]	0.09±0.02	0.27±0.02	0.50±0.04	**0.65±0.04**
	LD[27]	0.11±0.05	0.18±0.04	0.22±0.01	0.27±0.07
	DisCo[18]	0.07±0.00	0.26±0.00	0.48±0.00	0.53±0.00
	Dis-GU	0.13±0.09	**0.30±0.00**	**0.51±0.00**	0.55±0.04
Cars3D[23]	β-VAE[9]	0.08±0.02	0.14±0.02	0.79±0.08	0.98±0.01
	FactorVAE[1]	0.14±0.02	0.16±0.07	0.76±0.05	0.99±0.06
	DIP-VAE[26]	0.15±0.10	0.22±0.03	0.71±0.06	0.74±0.05
	CF[27]	0.07±0.04	0.24±0.03	0.81±0.04	0.93±0.04
	GS[27]	0.14±0.05	0.21±0.06	0.82±0.02	0.96±0.02
	DS[27]	0.12±0.06	0.23±0.03	0.74±0.04	0.98±0.03
	LD[27]	0.11±0.07	0.22±0.08	0.72±0.06	0.95±0.07
	DisCo[18]	0.06±0.00	0.19±0.00	0.88±0.00	**1.00±0.00**
	Dis-GU	**0.16±0.07**	**0.27±0.00**	**0.89±0.01**	0.98±0.03

注. 粗体表示最优值.

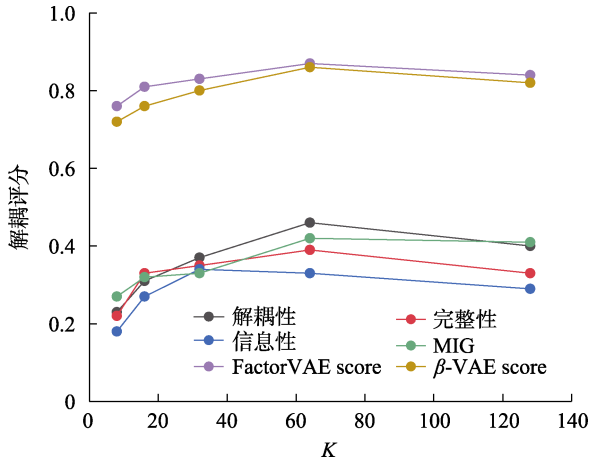

图 15　K 的参数分析

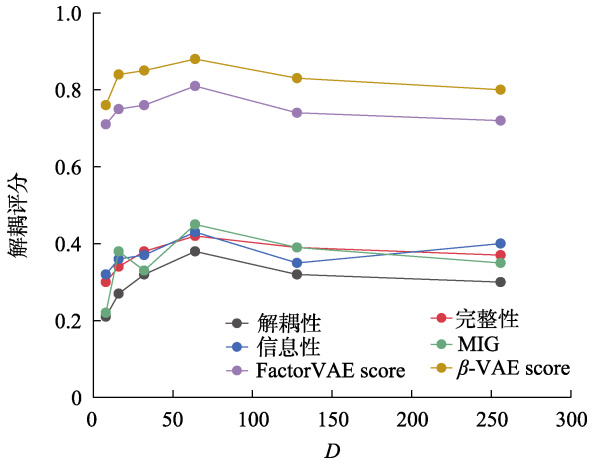

图 16　D 的参数分析

解耦性能最佳. 因此, 为了平衡推理成本和解耦性能, 在实验中设置超参数 D =64.

(3) 阈值 T . 在解耦表征学习阶段, 有许多在不同的方向上具有相同语义的样本, 为了识别这

些样本，在实验中设置阈值 T，并使这些样本与查询的相似度作为其伪标签. 表 2 所示为对 T 进行分析的结果. 可以看出，当 $T<0.95$ 时，Dis-GU 的解耦性能较差，可能会误判真实的样本，从而使对比损失的优化崩溃；当 $T\geqslant 0.95$ 时，Dis-GU 对 T 不敏感，当 $T=0.95$ 时，Dis-GU 的解耦性能最佳. 因此，根据经验设置，在实验中设置超参数 T=0.95.

表 2　T 的参数分析

T	解耦性	完整性	信息性	MIG	FactorVAE score	β-VAE score
0.80	0.28	0.16	0.39	0.20	0.87	0.90
0.85	0.31	0.18	0.37	0.21	0.89	0.91
0.90	0.33	0.23	0.35	0.22	0.91	0.93
0.95	**0.54**	**0.42**	0.40	**0.51**	**0.93**	**0.95**
0.98	0.51	0.41	**0.42**	0.49	0.92	0.93

注. 粗体表示最优值.

4　结　语

本文提出一种简单且有效的用于生成式视觉图像理解的解耦表征学习方法 Dis-GU. 该方法基于生成模型 Glow 的对比学习框架，通过构建可解释方向的发现和解耦表征学习这 2 大模块，实现联合学习并提取视觉图像表征的任务. 在 3 个数据集上与该领域多种典型的方法进行了实验对比，结果表明，Dis-GU 的解耦性能较优，获得的解耦表征学习结果易于解释.

Dis-GU 具有广泛的应用场景，尤其在图像生成的应用上，可同时解耦目标变化图像的多个潜在变化因子，提高样本的学习效率，并可控地生成丰富的图像. 但是，该方法也存在一定的局限性，未来将在 2 个方面进行深入研究，进一步提升模型的解耦性能：(1) 改进生成模型提取视觉图像特征的方式；(2) 提升模型的特征感知能力.

参考文献(References):

[1] Kim H, Mnih A. Disentangling by factorising[C] //Proceedings of the 35th International Conference on Machine Learning. New York: ICML, 2018: 2654-2663

[2] Lopez R, Regier J, Jordan M I, *et al*. Information constraints on auto-encoding variational Bayes[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2018: 6117-6128

[3] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4396-4405

[4] Chen R T Q, Li X C, Grosse R B, *et al*. Isolating sources of disentanglement in VAEs[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2018: 2615-2625

[5] Lin Z N, Thekumparampil K K, Fanti G, *et al*. InfoGAN-CR and modelcentrality: self-supervised model training and selection for disentangling GANs[C] //Proceedings of the 37th International Conference on Machine Learning. New York: ACM Press, 2020: Article No.569

[6] Li Zhixin, Zheng Yongzhe, Zhang Canlong, *et al*. Combining deep feature and multi-label classification for semantic image annotation[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(2): 318-326(in Chinese)
(李志欣, 郑永哲, 张灿龙, 等. 结合深度特征与多标记分类的图像语义标注[J]. 计算机辅助设计与图形学学报, 2018, 30(2): 318-326)

[7] Chen X, Duan Y, Houthooft R, *et al*. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM Press, 2016: 2180-2188

[8] Xiao T H, Hong J P, Ma J W. DNA-GAN: learning disentangled representations from multi-attribute images[OL]. [2024-01-02]. https://arxiv.org/abs/1711.05415

[9] Higgins I, Matthey L, Pal A, *et al*. Beta-VAE: learning basic visual concepts with a constrained variational framework[OL]. [2024-01-02]. https://openreview.net/forum?id=Sy2fzU9gl

[10] Yi Z L, Chen Z Q, Cai H, *et al*. BSD-GAN: branched generative adversarial network for scale-disentangled representation learning and image synthesis[J]. IEEE Transactions on Image Processing, 2020, 29: 9073-9083

[11] Shen Y J, Gu J J, Tang X O, *et al*. Interpreting the latent space of GANs for semantic face editing[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9240-9249

[12] Dalva Y, Pehlivan H, Hatipoglu O I, *et al*. Image-to-image translation with disentangled latent vectors for face editing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14777-14788

[13] Endo Y. User-controllable latent transformer for StyleGAN image layout editing[J]. Computer Graphics Forum, 2022, 41(7): 395-406

[14] Xu Y Y, Du Y, Xiao W P, *et al*. From continuity to editability: inverting GANs with consecutive images[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13890-13898

[15] Mo S T, Sun Z, Li C. Representation disentanglement in generative models with contrastive learning[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 1531-1540

[16] He K M, Fan H Q, Wu Y X, *et al*. Momentum contrast for un-

supervised visual representation learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9726-9735

[17] Le-Khac P H, Healy G, Smeaton A F. Contrastive representation learning: a framework and review[J]. IEEE Access, 2020, 8: 193907-193934

[18] Ren X C, Yang T, Wang Y W, *et al*. Learning disentangled representation by exploiting pretrained generative models: a contrastive learning view[OL]. [2024-01-02]. https://arxiv.org/abs/2102.10543

[19] Kingma D P, Dhariwal P. Glow: generative flow with invertible 1×1 convolutions[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM Press, 2018: 10236-10245

[20] Voynov A, Babenko A. Unsupervised discovery of interpretable directions in the GAN latent space[C] //Proceedings of the 37th International Conference on Machine Learning. New York: ICML, 2020: 9786-9796

[21] Wu Z R, Xiong Y J, Yu S X, *et al*. Unsupervised feature learning via non-parametric instance discrimination[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3733-3742

[22] Wen Zaidao, Wang Jiarui, Wang Xiaoxu, *et al*. A review of disentangled representation learning[J]. Acta Automatica Sinica, 2022, 48(2): 351-374(in Chinese)
(文载道, 王佳蕊, 王小旭, 等. 解耦表征学习综述[J]. 自动化学报, 2022, 48(2): 351-374)

[23] Reed S, Zhang Y, Zhang Y T, *et al*. Deep visual analogy-making[C] //Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM Press, 2015: 1252-1260

[24] Gondal M W, Wuthrich M, Miladinović Đ, *et al*. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM Press, 2019: Article No.1410

[25] Elizabeth Rani G, Sakthimohan M, Abhigna Reddy G, *et al*. MNIST handwritten digit recognition using machine learning[C] //Proceedings of the 2nd International Conference on Advance Computing and Innovative Technologies in Engineering. Los Alamitos: IEEE Computer Society Press, 2022: 768-772

[26] Kumar A, Sattigeri P, Balakrishnan A. Variational inference of disentangled latent concepts from unlabeled observations[OL]. [2024-01-02]. https://arxiv.org/abs/1711.00848

[27] Khrulkov V, Mirvakhabova L, Oseledets I, *et al*. Disentangled representations from non-disentangled models[OL]. [2024-01-02]. https://arxiv.org/abs/2102.06204